**A globally diverse reference alignment and panel for imputation of mitochondrial DNA variants**

Tim W McInerney1, Brian Fulton-Howard2, Christopher Patterson3,4, Devashi Paliwal1, Lars S Jermiin5,6,7,8, Hardip R Patel1, Judy Pa3,4, Russell H Swerdlow9, Alison Goate2, Simon Easteal1, Shea J Andrews2\*, for the Alzheimer’s Disease Neuroimaging Initiative†

1The John Curtin School of Medical Research, The Australian National University, Canberra, Australian Capital Territory, Australia

2Ronald M. Loeb Center for Alzheimer’s disease, Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York, USA

3Mark and Mary Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

4Department of Neurology, Alzheimer’s Disease Research Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

5CSIRO Land & Water, Commonwealth Scientific Industrial & Research Organization, Acton, ACT 2601, Australia

6Research School of Biology, Australian National University, Canberra, ACT 2601, Australia

7School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland

8Earth Institute, University College Dublin, Belfield, Dublin 4, Ireland

9Department of Neurology, Alzheimer’s Disease Center, University of Kansas, Fairway, KS, USA

\*Correspondence to: Shea Andrews, The Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Place, New York, NY 10029, USA.

Tel: +1-212-659-8632; E-mail: [shea.andrews@mssm.edu](mailto:shea.andrews@mssm.edu)

†Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at:

<http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf>

# **Abstract**

Motivation: Some existing mitochondrial DNA (mtDNA) datasets have been generated from genotyping microarrays or only the hypervariable regions, therefore they do not include full sequence information. For this reason they have limited capacity to assign mtDNA sequences to haplogroups (proxies for phylogenetic lineages). To address this limitation, we have created and curated a reference alignment and panel consisting of publicly available mtDNA sequences from which missing mtDNA single nucleotide variants can be imputed statistically. We have packaged these reference resources into a user-friendly imputation pipeline, MitoImpute.

Results: We downloaded and aligned publicly available complete human mtDNA sequences from GenBank. After filtering and quality control, we aligned 36,960 sequences, which was subsequently reformatted for use in IMPUTE2. We assessed the imputation accuracy of MitoImpute by measuring haplogroup and genotype concordance in data from the 1,000 Genomes Project and Alzheimer’s Disease Neuroimaging Initiative. The mean improvement of haplogroup assignment in the 1,000 Genomes samples was 42.7% (Matthew’s correlation coefficient of MCC=0.64). In ADNI cohort, we imputed missing SNVs. The results show that our reference alignment and panel can be used to impute missing mtSNVs in data obtained using microarrays. This improvement is particularly useful in longitudinal studies because it enables existing microarray data to be more accurately compared with more recent sequence data.

Availability

<https://github.com/sjfandrews/MitoImpute>

Supplementary Information.

Supplementary data are available at Bioinformatics online

## **Introduction**

Mitochondrial DNA (mtDNA) variation is informative about human evolution and can be associated with disease [(Gorman *et al.*, 2016)](https://paperpile.com/c/Ad0ZFB/PQ8Ck). mtDNA variation is often classified through the assignment of sequences to haplotype groups (haplogroups), which represent hypothetical points in the mtDNA phylogeny, with deeper points representing major haplogroups and thus proxies for evolutionary relationships. While whole-molecule sequencing of mtDNA is increasingly common, many datasets still only report mtDNA-derived single nucleotide variants (mtSNVs) obtained from microarrays. When different datasets report different mtSNVs, some SNVs will be present in some data sets and absent in others (and *vice versa*), so the datasets are neither directly comparable nor useful in combined analyses. Furthermore, microarrays that do not contain sufficient mtSNVs for haplogroup assignment will generate inaccurate estimates of human mtDNA variation. Imputing missing mtSNVs from a reference panel representative of human mtDNA diversity is required.

Several reference panels for imputing autosomal and sex-chromosomal variants have been created in recent years [(1000 Genomes Project Consortium *et al.*, 2015)](https://paperpile.com/c/Ad0ZFB/Vveu2); [(McCarthy *et al.*, 2016)](https://paperpile.com/c/Ad0ZFB/t2bf). However, to our knowledge, no large, globally diverse reference panel has been produced for mtDNA. The 1,000 Genomes Project dataset contains 2,504 mtDNA sequences, representing 26 populations, but many other populations (e.g., Pacific Islanders, Indigenous Australians, and Central Asians) are not represented. Furthermore, the panel derived from the 1000 Genomes Project dataset is not in a format readily available for imputation, adding further complexity to future analyses of these data. Samples from under-represented human populations are unlikely to have their haplotypes represented, even if they are common in their individual population, due to an ascertainment bias towards European and East Asian populations [(Popejoy and Fullerton, 2016)](https://paperpile.com/c/Ad0ZFB/Rrk5K). This represents a significant problem for both medical and evolutionary research, as historical datasets from these populations would need to have missing mtSNVs imputed from other populations that do not include haplotypes present in the sample panel populations. Inability to capture haplotypes with a disease-causing mtSNV will lead to a failure to understand the genetic basis behind human disease [(Popejoy and Fullerton, 2016; Sirugo, Williams and Tishkoff, 2019)](https://paperpile.com/c/Ad0ZFB/Rrk5K+UAa8n). Finally, in some instances, where mtDNA reference panels have been created, they have not been made public [(Hudson *et al.*, 2014; Gonçalves *et al.*, 2018)](https://paperpile.com/c/Ad0ZFB/aZvpe+kR73). This again presents a hurdle for users without experience in multiple sequence alignment (MSA).

To ensure that allelic frequencies, and therefore probabilities of observing a given haplotype, are accurately estimated, it is pertinent that homologous sites in different mtDNA genomes are aligned correctly. Typically, this entails inserting alignment gaps (‘–’) between some of the nucleotides in some of the sequences in an MSA. When a site in an MSA contains one or several alignment gaps, we can infer that insertions and/or deletions (jointly labelled indels) must have occurred during the evolution of the genomes in the MSA. However, placing alignment gaps correctly in an MSA turns out to be more challenging than most scientists realise—in many cases, MSA methods prioritize aligning the alignment gaps over aligning the nucleotides [(Golubchik *et al.*, 2007)](https://paperpile.com/c/Ad0ZFB/dXHy). Obviously, that does not reflect biological reality. Experienced bioinformaticians may manually curate their MSAs to correctly place the alignment gaps [(Morrison, 2009, 2015)](https://paperpile.com/c/Ad0ZFB/4Pzu+fbx8); however, many others do not have the expertise, time, and/or resources to do so (see, e.g., the alignment of flavivirus genomes used by Lessler et al. 2016). To enable future population genetic and disease-related research that relies on accurate identification of human mtDNA haplotypes, it is clear that a publicly available high-quality MSA, comprising globally diverse sequences, and reference panel will be needed.

Given the lack of these vital resources, we have created a large (*n*=36,960) globally diverse MSA using automated alignment software and manual curation by experienced researchers. We then converted this MSA into a reference panel and have made it publicly available through GitHub. We also have developed a SnakeMake pipeline (MitoImpute) for easy imputation of mtSNVs through the IMPUTE2 framework. We validated the performance of MitoImpute using *in silico* microarrays (ISMs) derived from 1000 Genomes Project’s [(1000 Genomes Project Consortium *et al.*, 2015)](https://paperpile.com/c/Ad0ZFB/Vveu2) whole-genome sequence (WGS) data, and empirical data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [(Saykin *et al.*, 2010)](https://paperpile.com/c/Ad0ZFB/s1iqn).

## **Methods**

### Reference Alignment and Reference Panel

Whole human mtDNA sequences were downloaded from GenBank on 2018-07-18 by adapting the MitoMap [(Lott et al., 2013)](https://paperpile.com/c/Ad0ZFB/mKM50) search term (Supplementary Methods). This returned 44,299 complete human mtDNA sequences and excluded archaic and ancient sequences. Sequences were aligned to an existing reference alignment (unpublished, supplementary) in batches of 2,500 using MAFFT [(Katoh and Standley, 2013)](https://paperpile.com/c/Ad0ZFB/Sv4E1) in Geneious v10.2.6 [(Kearse et al., 2012)](https://paperpile.com/c/Ad0ZFB/KUlom). The existing reference alignment used included the revised Cambridge Reference Sequence (rCRS) [(Andrews et al., 1999)](https://paperpile.com/c/Ad0ZFB/RL6RB) so the standardised site numbering convention is maintained. Likewise, this new Reference MSA contains the rCRS as the first sequence. As such, sites that introduced gaps in the reference alignment were removed to maintain consistent rCRS nucleotide position numbering.

To improve the quality of the Reference Panel, sequences with ≥5 ambiguous characters or ≥8 gaps were removed from the alignment. This threshold was set to enable inclusion of haplogroup B sequences which averaged 7 gaps relative to other the sequences. Using this quality filter led to a Reference Panel of 36,960 sequences (Supplementary Table 1).

AliStat v1.11 [(Wong *et al.*, 2020)](https://paperpile.com/c/Ad0ZFB/T55k) was used to quantify the completeness of the new Reference MSA. The Reference Panel was created by converting the Reference MSA to formats compatible with IMPUTE2 [reference needed].

### **Validation Panel**

The ISMs were obtained by sub-setting mtSNVs present in 1000 Genomes Project Phase 3 WGS data (*n*=2,535) to those from existing commercially available microarrays. Microarray information was obtained from strand orientation files available from the Wellcome Centre (<http://www.well.ox.ac.uk/~wrayner/strand/>), with 101 strand files containing mtSNVs (Supplementary Table 2). Haplogroup assignment for the WGS data and the ISMs was performed using Hi-MC [(Smieszek *et al.*, 2018)](https://paperpile.com/c/Ad0ZFB/49KJC).

### **Imputation**

We used the IMPUTE2 chromosome X protocol for imputation [(Howie, Donnelly and Marchini, 2009; Gonçalves *et al.*, 2018)](https://paperpile.com/c/Ad0ZFB/PumIw+aZvpe). No recombination was assumed; therefore, we applied a uniform recombination rate of *r*=0 across all sites. The Markov chain Monte Carlo step in IMPUTE2 was used to account for phase uncertainty in recombining diploid data [(Howie, Donnelly and Marchini, 2009)](https://paperpile.com/c/Ad0ZFB/PumIw) but we did not perform this step as our data is non-recombining and haploid.

The effect of varying the number of sequences in the reference alignment (khap) was estimated by setting khap to 100, 250, 500, 1,000, 2,500, 5,000, 10,000, 20,000, and 30,000. We tested the ability of our pipeline to impute rare variants by filtering the Reference Panel to exclude variants with minor allele frequencies (MAF) of 1%, 0.5% and 0.1%, resulting 409, 682 and 1874 mtSNVs, respectively (Supplementary Tables 3). Imputation accuracy was assessed using Matthew’s (1975) Correlation Coefficient (MCC) for genotype concordance and Hi-MC [(Smieszek *et al.*, 2018)](https://paperpile.com/c/Ad0ZFB/49KJC) for haplogroup assignment, with the WGS data used as the truth set. Linear mixed-model ANOVA was used to assess the meaningful difference in haplogroup assignment and MCC (mean of mtSNVs per ISM) for different parameters tested for khap and MAF. Pipelines for implementing our imputation protocol and reproducing our results were created in SnakeMake [(Köster and Rahmann, 2012)](https://paperpile.com/c/Ad0ZFB/g375i).

## **Results**

### **Reference Alignment and Reference Panel**

To comply with minimum reporting standards for MSAs, completeness metrics of the Reference MSA were computed (Table 1). As described in Wong et al. (2020), *Ca* is the completeness of the alignment, *Cr* is the completeness of the *r*th sequence, *Cc* is the completeness of the *c*th site, and *Cij* is the completeness of the *i*th and *j*th sequences. Overall, the Reference MSA is highly complete (*Ca* > 0.99). Individual sequences are also largely complete (*Cr*), with the least complete sequence containing completely-specified nucleotide at 91% of its sites and the most complete sequence containing completely-specified nucleotides at all of it sites. The least complete site in the MSA contained completely-specified nucleotides in 44.3% of sequences, and the most complete sites had completely-specified nucleotides in all of the sequences. The proportion of homologous sites with completely-specified nucleotides at in both sequences (*Cij*) ranged from 83% so 100%, suggesting that the majority of sequence pairs contain enough information to quantify evolutionary distances. Sites and sequences missing a substantial number of nucleotide states were removed in the filtration processes as described in the Methods section.

|  |  |
| --- | --- |
| Table 1. Completeness metrics for the Reference MSA, obtained using AliStat. | |
| Feature | Value(s) |
| Sequences | 44,299 |
| Sites | 16,569 |
| Completeness Score (*Ca*) | 0.9997 |
| *C*-score for individual sequences (*Cr*) [min-max] | 0.9119 - 1.0000 |
| *C*-score for individual sites (*Cc*) [min-max] | 0.4429 - 1.0000 |
| *C*-score for pairs of sequences (*Cij*, *i*≠*j*) [min-max] | 0.8314 - 1.0000 |

The Reference Panel consisted of 7,128 (19.3%) sequences for which GenBank metadata on geographic provenance was available (Table 1) representing 49 countries. Within-country provenance was available for 1,167 samples, taking the total number of identified regions of provenance to 103. This included smaller ethnic groups such as Yami Taiwanese, Morrocan Berbers, Pacific Islanders, Indigenous Australians, and people from Central Asia and Siberia. There is, however, still a distinct bias towards European (3,855; 54.1% of sequences with provenance) and East Asian (2,065; 29.0%) samples. All major haplogroups are represented in the Reference Panel (Table 1), including rare haplogroups such as haplogroup S which is endemic to Indigenous Australians, haplogroup L5 which is found in Mbuti Pygmies, haplogroup L6 which is found in low frequencies in Yemen and Ethiopia, and haplogroups O and Q which are found exclusively in the Pacific Islands. Haplogroup B was the most filtered haplogroup after quality control on the Reference Panel, removing 3,395 sequences (or, 46% of removed sequences), leaving only 273 haplogroup B sequences. Haplogroup H was also heavily filtered following QC (1,376; 19%), however haplogroup H was still highly represented in the final reference panel (n=7,644). All other haplogroups had only a small fraction of their sequences removed during QC.

### ***In silico* Microarrays**

#### ***Parameter Tuning***

When compared to non-imputed data, haplogroup concordance improved by 42.7%, 44.6%, and 43.3% for MAF = 1%, 0.5%, and 0.1%, respectively (Supplementary Table 4; Supplementary Figure 1). Variation in this success rate was within the expected range (AVOVA, *p*=0.6). For genotype concordance, the best results were obtained for MAF = 1%; here the variation in success rate was significant (ANOVA, *p*<0.0001, Supplementary Table 5; Supplementary Figure 2). The number of reference haplotypes used had a noticeable effect on haplogroup and genotype concordance (ANOVA, *p*<0.0001, Supplementary Table 6; Supplementary Figure 3, 4). There was no significant difference between the top four khap parameter settings (khap = 100, 250, 500, 1000). Larger khap parameter settings performed comparatively poorly, indicating a reduced ability to correctly assign haplogroups for some ISMs.

#### ***Overall Microarray Performance***

Using our recommended settings (khap = 500, MAF = 1%), the average haplogroup assignment accuracy was 89.3% (95% Confidence Interval [CI] = 87.4, 91.2) following imputation, an increase of 42.7% (95% CI = 40.1%, 45.23%) (Supplementary Table 7). The best-performing ISM (Illumina HumanHap 240S) correctly assigned 99.8% of haplogroups after haplotype imputation, a small improvement of 0.8%. The worst performing group of ISMs (HumanOmni1-Quads) correctly assigned 52.3% of haplogroups after imputation compared to 12.9% before imputation. Correct assignment for the worst performing individual ISM (HumanOmni 2.5) increased from 4.9% to 64.0% after imputation. The greatest improvement was 64.8% for the HumanCore ISMs. In terms of genotype concordance, the mean MCC = 0.64 (95% CI = 0.60, 0.68, Supplementary Table 7), to MCC = 0.97 for the best performing ISM (Infinium Global Screening Array-24v2) and to MCC = 0.10 for the worst performing ISM (HumanOmni 2.5).

#### ***Overall Haplogroup Concordance***

Concordance of individual haplogroups was estimated at the macro-haplogroup level. Prior to imputation, less than 49% of sequences from haplogroups M, HV, D, L, A, H, J, W, I, V were assigned to their correct haplogroup (Supplementary Table 8). Imputation improved haplogroup assignment by between 30% and 83%. Microarray assignment was relatively good (>74%) for haplogroups R, B, U, N, C, T, K, so improvement from imputation was, correspondingly, minor to moderate (0.1%-18%). Haplogroups JT and X showed no improvement.

### **Alzheimer’s Disease Neuroimaging Initiative**

We applied MitoImpute to data from 258 participants in the ADNI study who had provided both WGS [(Ridge *et al.*, 2018)](https://paperpile.com/c/Ad0ZFB/LtOdg) and genotyping data [(Saykin *et al.*, 2010)](https://paperpile.com/c/Ad0ZFB/s1iqn) (Supplementary Table 9). The ADNI genotype data were mapped to the rCRS. Hi-MC [(Smieszek *et al.*, 2018)](https://paperpile.com/c/Ad0ZFB/49KJC) was used to assign haplogroups to the WGS, genotyped, and imputed data. Genotype data assigned the correct haplogroup to 31.4% of samples, which improved to 91.9% (Supplementary Table 10) after imputation. The corresponding improvement for macro-haplogroups was 37.2% to 95%. Eight of nineteen macro-haplogroups showed no improvement as the genotype data provided perfect or near-perfect haplogroup assignment. Haplogroups J, L2, M, V, W, X all improved from 0% to 100% correct assignment. Haplogroup H was the most frequently observed and showed an improvement of 5.8% to 100%. Haplogroups N & R were the worst performing post-imputation at 25% and 36.4%, respectively (Supplementary Table 11). Following imputation, the mean genotype concordance per mtSNV was MCC = 0.71 (95% CI = 0.66, 0.75).

## **Discussion**

Investigations into the genetic basis of human mitochondrial disease and of evolutionary history are reliant upon the accurate alignment of homologous nucleotide positions, and complete mtDNA sequences [(Kumar and Filipski, 2007)](https://paperpile.com/c/Ad0ZFB/F0Am). These two factors, in turn, benefit from globally diverse sequences being included in MSAs used in these investigations. Datasets of incomplete mtSNVs can be mitigated by imputation of missing variants, however accurate alignment of sequences and consistent placement of gap character states is fraught with difficulty and time consuming for even experienced bioinformaticians [(Golubchik *et al.*, 2007)](https://paperpile.com/c/Ad0ZFB/dXHy). Lack of publicly available reference MSAs and reference panels, therefore, presents a limitation to researchers investigating mitochondrial disease or evolutionary history. We address this limitation by creating a reference MSA from 36,960 globally diverse mtDNA sequences, which was manually curated by experienced researchers to ensure consistency of the placement of gap character states. Aligning novel sequences to our reference alignment will alleviate the pressures of the alignment process by providing a guide for these new sequences. Additionally, as a curated MSA, our reference MSA can be subsampled for use in answering evolutionary and disease-associated research questions. Furthermore, the reference MSA can be used as a reference panel for the imputation of mtSNVs. To our best knowledge, this is the largest and most genetically and geographically diverse curated mtDNA reference panel publicly available; this reference panel should facilitate comparison and combined analyses across datasets of differing age and completeness. The reference panel has been packaged into a user-friendly mtSNV imputation pipeline, MitoImpute.

We evaluated how accurately we could impute mtSNVs using our reference panel, as measured by concordance of assigned haplogroups and Matthew’s correlation coefficient of genotypes. Across most ISMs we were able to improve haplogroup assignment, suggesting we are successfully imputing phylogenetically informative mtSNVs. As all haplogroups, except for haplogroups JT and X, experienced an average improvement >30%, this suggests that the reference panel is not biased towards improvement for certain lineages over others. The addition of new sequences to the reference panel will only further increase accurate haplogroup assignment in populations or mtDNA lineages that are still underrepresented. We also tested the practical usage of our reference panel by imputing mtSNVs in the ADNI dataset, demonstrating that the reference panel and imputation pipeline dramatically increased the correct haplogroup assignment. Given that there are 499 samples in the ADNI genotyping dataset that were not resequenced in subsequent phases, this demonstrates the utility of our reference panel for long-term studies that need to bring their older, incomplete dataset to the same standard as newer, complete datasets.

Performance testing of the MitoImpute pipeline using ISM revealed some counterintuitive results. These tests included different parameter values on the reference panel itself and parameter values of the imputation protocol. For instance, one would expect that including rarer haplotypes into the reference panel by decreasing the MAF threshold would lead to increases in imputation accuracy [(Sariya *et al.*, 2019](https://paperpile.com/c/Ad0ZFB/zTpe); [Huang *et al.*, 2015](https://paperpile.com/c/Ad0ZFB/1e60); [Das *et al.*, 2016](https://paperpile.com/c/Ad0ZFB/3CV5); [Zheng *et al.*, 2012](https://paperpile.com/c/Ad0ZFB/xF88); [Browning and Browning, 2009)](https://paperpile.com/c/Ad0ZFB/UzxL). However, the best performing of the reference panel parameter settings was the highest MAF threshold (MAF ≥ 1%; Supplemental Figure 2). This problem may be mitigated by partitioning the reference panel into sequences from geographic regions or populations similar to study samples, rather than using the ‘global’ reference panel [(Mitt *et al.*, 2017](https://paperpile.com/c/Ad0ZFB/WANm); [Surakka *et al.*, 2016)](https://paperpile.com/c/Ad0ZFB/lx58). This result warrants further investigation. We suspect the decrease in imputation accuracy is due ISMs with few mtSNVs being unable to ‘decide’ which reference haplotype to impute from, in some cases making an erroneous decision. Another seemingly counterintuitive result is the decrease in imputation accuracy as the khap parameter increases. Increasing the khap parameter increases the number of haplotypes in the reference panel from which IMPUTE2 will impute. We suspect that increasing the number of reference haplotypes beyond 1,000 leads to a greater chance of mismatch between the incomplete sample haplotypes and the reference panel haplotypes, particularly in ISMs with few mtSNVs. The limitations of the MAF and khap parameters, we suspect, is due to a dearth of mtSNVs in some ISMs. Datasets with a small number of variants from which to impute missing mtSNVs will always present this limitation, and we recommend users must proceed with caution when using these datasets for subsequent analyses.

Our reference panel provides an opportunity for datasets with limited mitochondrial genetic variation to be analyzed with a more complete set of genetic variants and a more accurate assignment of haplogroups. The global disparity in medical research is evident in the high proportion of European individuals (~78%) association study catalogues [(Sirugo, Williams and Tishkoff, 2019)](https://paperpile.com/c/Ad0ZFB/UAa8n). The 1,000 Genomes Project phase 3 includes 2,504 individuals from 26 populations, however these individuals were often sampled from 1-3 cities within geographically diverse countries, such as China. Our reference panel contains sequences from at least 103 regions in at least 49 countries, capturing a more globally-representative sample of mitochondrial genetic diversity. The diversity included in our reference panel will allow researchers to perform imputation in under-represented human populations, contributing to solving the disparity in medical genomic research.

# **Acknowledgments**

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer’s Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

# **Funding**

Dr. Judy Pa and Christopher Patterson were supported by the National Institute on Aging (R01AG054617 PI: Judy Pa). SJA, BFH and AMG are supported by the JPB Foundation (<http://www.jpbfoundation.org>). RHS is supported by P30 AG035982.

# **Conflicts of interest**

AMG served on the scientific advisory board for Denali Therapeutics from 2015-2018. She has also served as a consultant for Biogen, AbbVie, Pfizer, GSK, Eisai and Illumina.

## **References**

[1000 Genomes Project Consortium *et al.* (2015) ‘A global reference for human genetic variation’, *Nature*, 526(7571), pp. 68–74.](http://paperpile.com/b/Ad0ZFB/Vveu2)

[Andrews, R. M. *et al.* (1999) ‘Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA’, *Nature genetics*, 23(2), p. 147.](http://paperpile.com/b/Ad0ZFB/RL6RB)

[Browning, B. L. and Browning, S. R. (2009) ‘A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals’, *American journal of human genetics*, 84(2), pp. 210–223.](http://paperpile.com/b/Ad0ZFB/UzxL)

[Das, S. *et al.* (2016) ‘Next-generation genotype imputation service and methods’, *Nature genetics*, 48(10), pp. 1284–1287.](http://paperpile.com/b/Ad0ZFB/3CV5)

[Golubchik, T. *et al.* (2007) ‘Mind the gaps: evidence of bias in estimates of multiple sequence alignments’, *Molecular biology and evolution*, 24(11), pp. 2433–2442.](http://paperpile.com/b/Ad0ZFB/dXHy)

[Gonçalves, V. F. *et al.* (2018) ‘Examining the role of common and rare mitochondrial variants in schizophrenia’, *PloS one*, 13(1), p. e0191153.](http://paperpile.com/b/Ad0ZFB/aZvpe)

[Gorman, G. S. *et al.* (2016) ‘Mitochondrial diseases’, *Nature Reviews Disease Primers*, 2, p. 16080.](http://paperpile.com/b/Ad0ZFB/PQ8Ck)

[Howie, B. N., Donnelly, P. and Marchini, J. (2009) ‘A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies’. Edited by N. J. Schork. doi:](http://paperpile.com/b/Ad0ZFB/PumIw) [10.1371/journal.pgen.1000529](http://dx.doi.org/10.1371/journal.pgen.1000529)[.](http://paperpile.com/b/Ad0ZFB/PumIw)

[Huang, J. *et al.* (2015) ‘Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel’, *Nature communications*, 6, p. 8111.](http://paperpile.com/b/Ad0ZFB/1e60)

[Hudson, G. *et al.* (2014) ‘Recent mitochondrial DNA mutations increase the risk of developing common late-onset human diseases’, *PLoS genetics*, 10(5), p. e1004369.](http://paperpile.com/b/Ad0ZFB/kR73)

[Katoh, K. and Standley, D. M. (2013) ‘MAFFT multiple sequence alignment software version 7: improvements in performance and usability’, *Molecular biology and evolution*, 30(4), pp. 772–780.](http://paperpile.com/b/Ad0ZFB/Sv4E1)

[Kearse, M. *et al.* (2012) ‘Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data’, *Bioinformatics* , 28(12), pp. 1647–1649.](http://paperpile.com/b/Ad0ZFB/KUlom)

[Köster, J. and Rahmann, S. (2012) ‘Snakemake--a scalable bioinformatics workflow engine’, *Bioinformatics* , 28(19), pp. 2520–2522.](http://paperpile.com/b/Ad0ZFB/g375i)

[Kumar, S. and Filipski, A. (2007) ‘Multiple sequence alignment: in pursuit of homologous DNA positions’, *Genome research*, 17(2), pp. 127–135.](http://paperpile.com/b/Ad0ZFB/F0Am)

[Lott, M. T. *et al.* (2013) ‘mtDNA Variation and Analysis Using Mitomap and Mitomaster’, *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, 44, pp. 1.23.1–26.](http://paperpile.com/b/Ad0ZFB/mKM50)

[Matthews, B. W. (1975) ‘Comparison of the predicted and observed secondary structure of T4 phage lysozyme’, *Biochimica et biophysica acta*, 405(2), pp. 442–451.](http://paperpile.com/b/Ad0ZFB/oPpyw)

[McCarthy, S. *et al.* (2016) ‘A reference panel of 64,976 haplotypes for genotype imputation’, *Nature genetics*, 48(10), pp. 1279–1283.](http://paperpile.com/b/Ad0ZFB/t2bf)

[Mitt, M. *et al.* (2017) ‘Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage WGS-based imputation reference panel’, *European journal of human genetics: EJHG*, 25(7), pp. 869–876.](http://paperpile.com/b/Ad0ZFB/WANm)

[Morrison, D. A. (2009) ‘Why would phylogeneticists ignore computerized sequence alignment?’, *Systematic biology*, 58(1), pp. 150–158.](http://paperpile.com/b/Ad0ZFB/4Pzu)

[Morrison, D. A. (2015) ‘Is Sequence Alignment an Art or a Science?’, *Systematic botany*, 40(1), pp. 14–26.](http://paperpile.com/b/Ad0ZFB/fbx8)

[van Oven, M. and Kayser, M. (2009) ‘Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation’, *Human mutation*, 30(2), pp. E386–E394.](http://paperpile.com/b/Ad0ZFB/eACa9)

[Popejoy, A. B. and Fullerton, S. M. (2016) ‘Genomics is failing on diversity’, *Nature*, pp. 161–164.](http://paperpile.com/b/Ad0ZFB/Rrk5K)

[Ridge, P. G. *et al.* (2018) ‘Assembly of 809 whole mitochondrial genomes with clinical, imaging, and fluid biomarker phenotyping’, *Alzheimer’s & dementia: the journal of the Alzheimer's Association*, 14(4), pp. 514–519.](http://paperpile.com/b/Ad0ZFB/LtOdg)

[Sariya, S. *et al.* (2019) ‘Rare Variants Imputation in Admixed Populations: Comparison Across Reference Panels and Bioinformatics Tools’, *Frontiers in genetics*, 10, p. 239.](http://paperpile.com/b/Ad0ZFB/zTpe)

[Saykin, A. J. *et al.* (2010) ‘Alzheimer’s Disease Neuroimaging Initiative biomarkers as quantitative phenotypes: Genetics core aims, progress, and plans’, *Alzheimer’s & dementia: the journal of the Alzheimer's Association*, 6(3), pp. 265–273.](http://paperpile.com/b/Ad0ZFB/s1iqn)

[Sirugo, G., Williams, S. M. and Tishkoff, S. A. (2019) ‘The Missing Diversity in Human Genetic Studies’, *Cell*, 177(1), pp. 26–31.](http://paperpile.com/b/Ad0ZFB/UAa8n)

[Smieszek, S. *et al.* (2018) ‘Hi-MC: a novel method for high-throughput mitochondrial haplogroup classification’. doi:](http://paperpile.com/b/Ad0ZFB/49KJC) [10.7717/peerj.5149](http://dx.doi.org/10.7717/peerj.5149)[.](http://paperpile.com/b/Ad0ZFB/49KJC)

[Surakka, I. *et al.* (2016) ‘The rate of false polymorphisms introduced when imputing genotypes from global imputation panels’, *Genetics*. bioRxiv.](http://paperpile.com/b/Ad0ZFB/lx58)

[Wong, T. K. F. *et al.* (2020) ‘A minimum reporting standard for multiple sequence alignments’, *bioRxiv*. doi:](http://paperpile.com/b/Ad0ZFB/T55k) [10.1101/2020.01.15.907733](http://dx.doi.org/10.1101/2020.01.15.907733)[.](http://paperpile.com/b/Ad0ZFB/T55k)

[Zheng, H.-F. *et al.* (2012) ‘Effect of genome-wide genotyping and reference panels on rare variants imputation’, *Journal of genetics and genomics = Yi chuan xue bao*, 39(10), pp. 545–550.](http://paperpile.com/b/Ad0ZFB/xF88)

## **Supplementary Information**

**Supplementary Methods**

The following search term was used to identify whole human mtDNA sequences from GenBank on 2018-07-18:

(016500[SLEN]:016600[SLEN]) AND Homo[Organism] AND mitochondrion[FILT] AND complete genome NOT (Homo sp. Altai OR Denisova hominin OR neanderthalensis OR heidelbergensis OR consensus OR ancient human remains OR shotgun)

## **Supplementary Tables**

Please find all tables at the following [link](https://docs.google.com/spreadsheets/d/1Km7TI6vZeAgH10Gye6SwFALLsd_WQB2UGWGbMFDA4Bk/edit?usp=sharing):